

Node Mapping and the Parallel Performance of LS3DF code on BG/P

Zhengji Zhao¹⁾ and Lin-Wang Wang²⁾

1) National Energy Research Scientific Computing Center (NERSC)

2) Computational Research Division (CRD)

Lawrence Berkeley National Laboratory

(LS3DF: Linearly Scaling 3 Dimensional Fragment)



ScicomP 10



San Francisco, CA



Quantum mechanical calculations are computationally expensive

$$\left[-\frac{1}{2}\nabla^2 + V_{tot}(r)\right]\psi_i(r) = \varepsilon_i\psi_i(r)$$

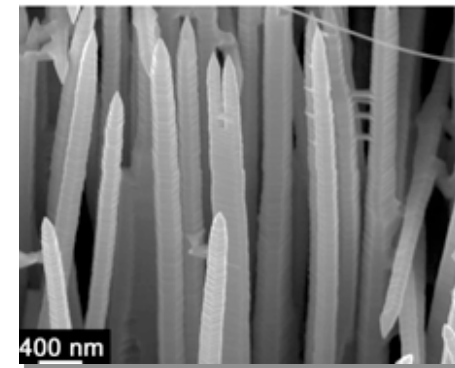
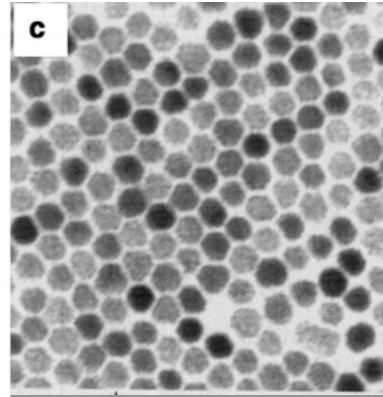
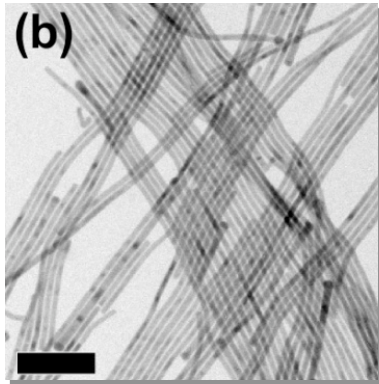
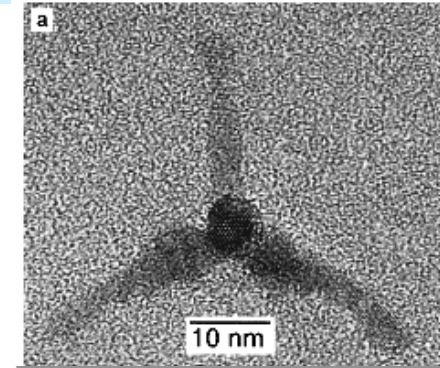
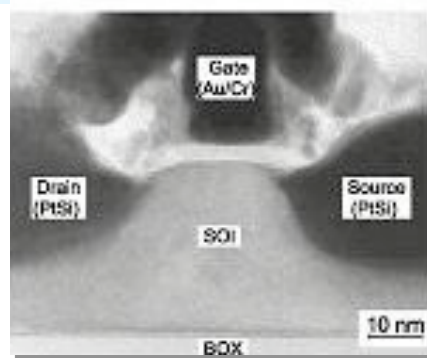
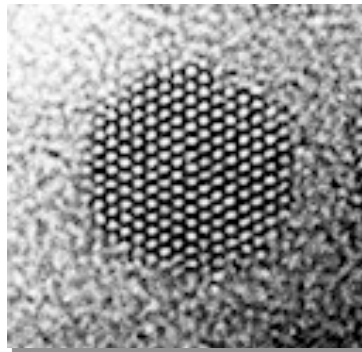
$$\int \psi_i^*(r)\psi_j(r)d^3r = \delta_{ij}, \quad i=1,\dots,M$$

- ❖ If the size of the system is N :
- ❖ N coefficients to describe one wavefunction $\psi_i(r)$
- ❖ $i = 1, \dots, M$ wavefunctions $\psi_i(r)$, M is proportional to N .
- ❖ Orthogonalization: $\int \psi_i^*(r)\psi_j(r)d^3r = \delta_{ij}$, algorithm $\propto N^*M^2$ floating point operations, *i.e.*, N^3 scaling.

This non-linear eigenvalue problem is solved iteratively. The repeated calculations of these orthogonal wavefunctions make the computation expensive, $O(N^3)$.



Nanostructures have wide applications including: solar cells, biological tags, electronics devices



- ❖ Different electronic structures than bulk materials
- ❖ 1,000 ~ 100,000 atom systems are too large for direct $O(N^3)$ *ab initio* calculations
- ❖ $O(N)$ computational methods are required
- ❖ Parallel supercomputers critical for the solution of these systems



Linearly Scaling 3 Dimensional Fragment method (LS3DF)

- ❖ Ab initio electronic structure code for large systems
- ❖ $O(N)$ code
- ❖ Divide and conquer method
 - A novel divide and conquer scheme with a new approach for patching the fragments together
 - Uses overlapping positive and negative fragments
 - Minimizes artificial boundary effects
- ❖ Young code
 - Started with several component codes connected with a shell script in 2006
 - Scaled up to thousands of processors in 2008 (Gordon Bell Prize 2008 for algorithmic innovation)
 - Has been used to calculate many nano systems together with other codes (the code is in continuous development)



Flops achieved on Franklin, Intrepid, and Jaguar



- ❖ 135 Tflops/s on 36,864 processors of the quad-core Cray XT4 Franklin at NERSC, 40% efficiency



- ❖ 224 Tflops/s on 163,840 processors of the BlueGene/P Intrepid at ALCF, 40% efficiency

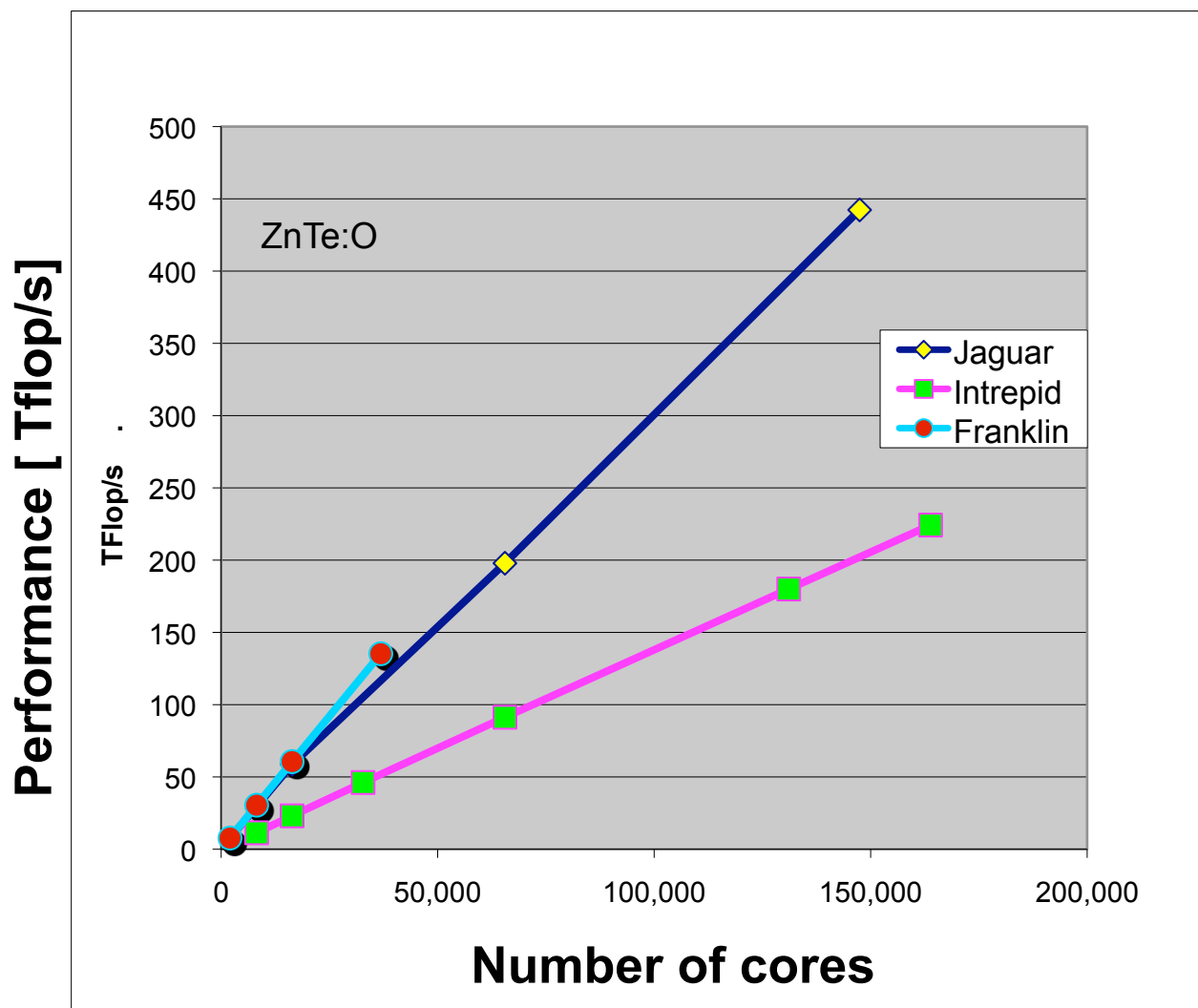


- ❖ 442 Tflops/s on 147,456 processors of the Cray XT5 Jaguar at NCCS, 33% efficiency

**For the largest physical system (36,000 atoms),
LS3DF is 1000 times faster than direct DFT codes**



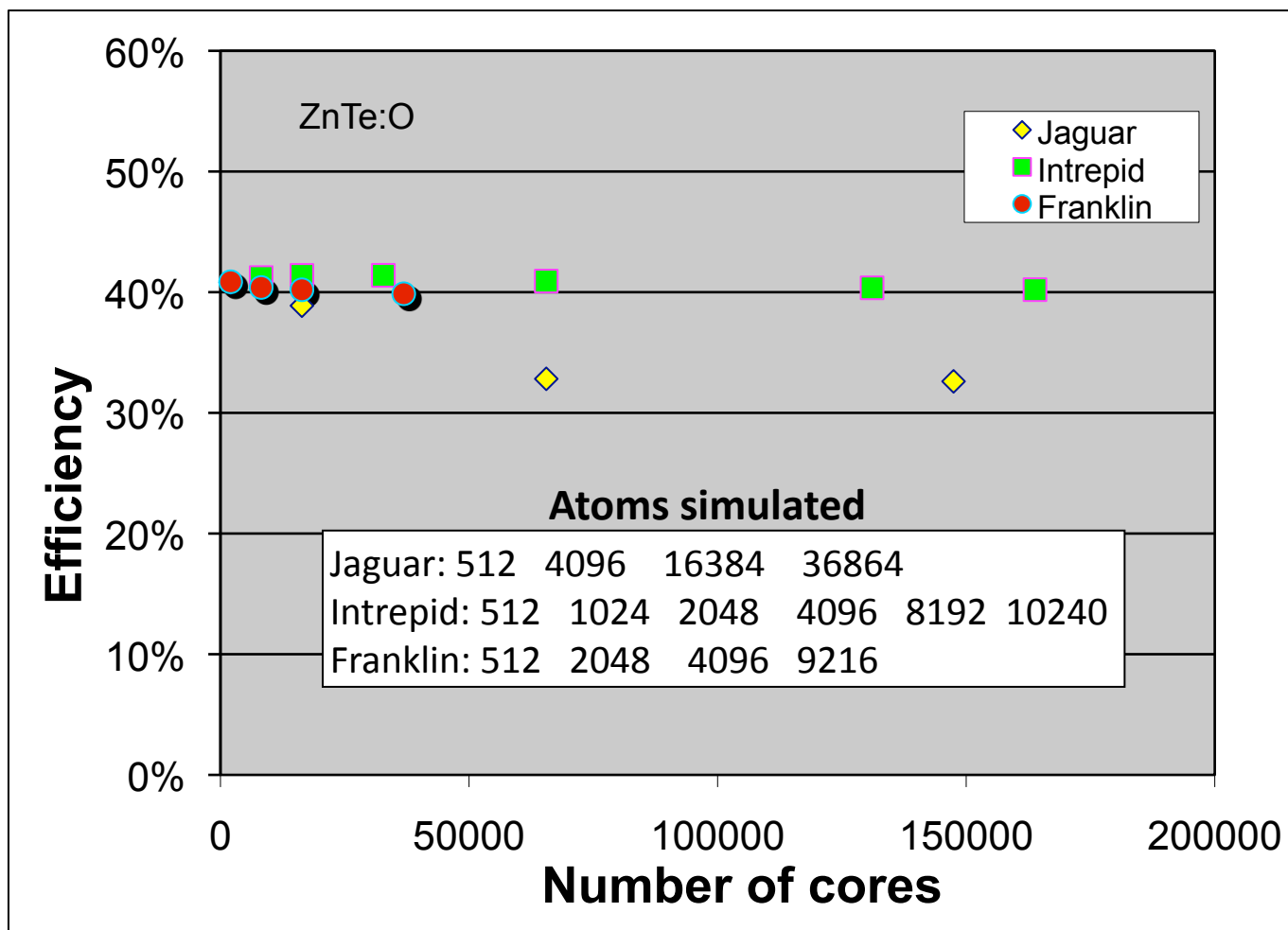
Weak scaling of LS3DF



Note: Ecut = 60Ryd with *d* states, up to 36864 atoms



Weak scaling of LS3DF

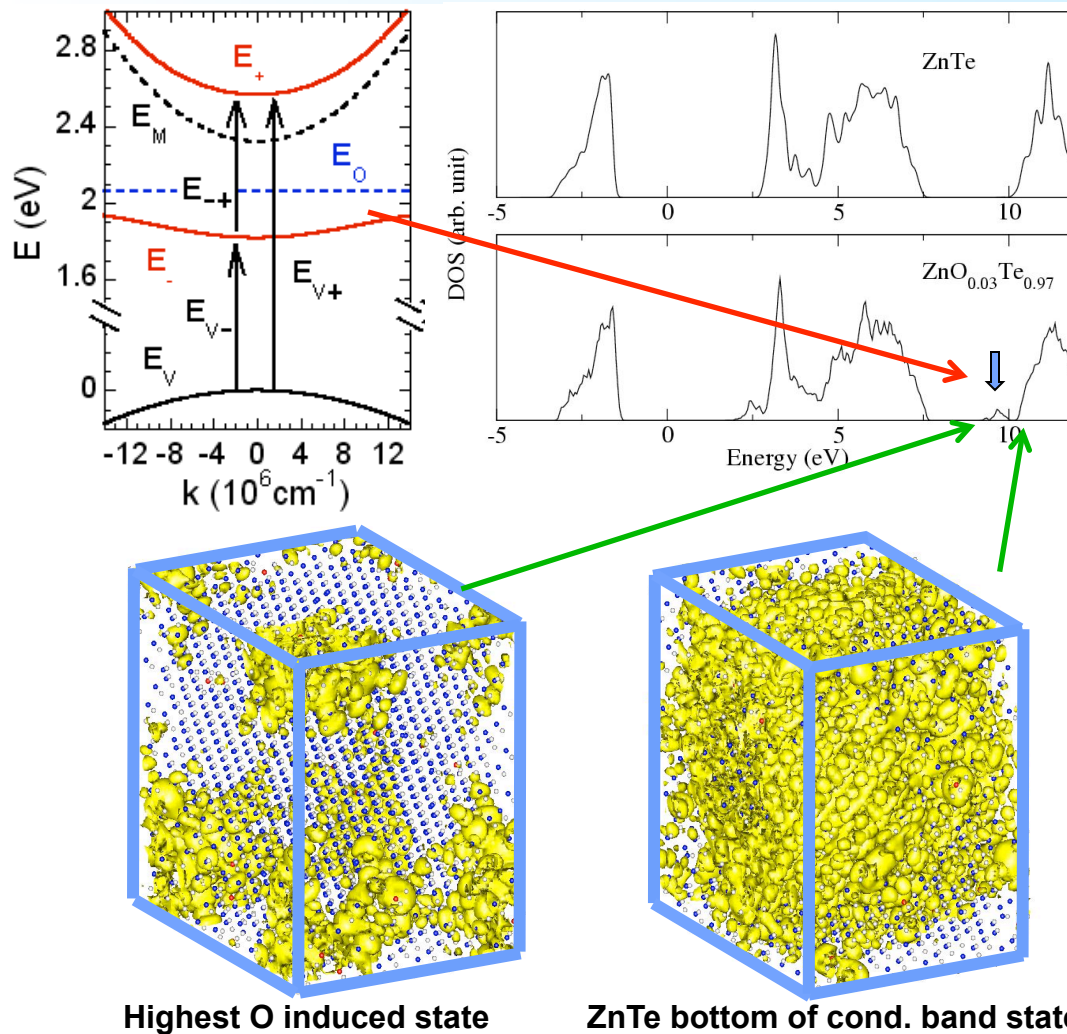


Note: Ecut = 60Ryd with *d* states, up to 36864 atoms

ScicomP 16, May 10-14, 2010, San Francisco, CA



Can one use an intermediate state to improve solar cell efficiency?



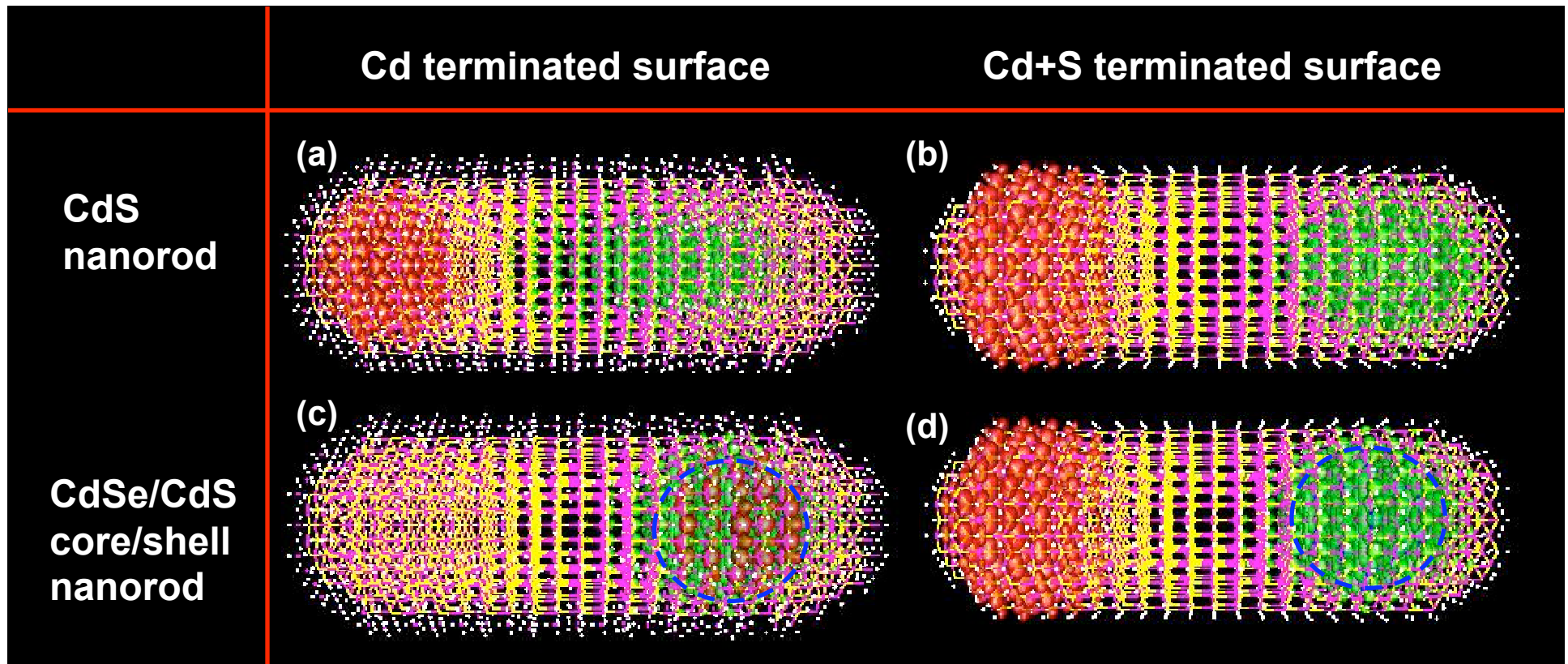
- ❖ Single band material theoretical PV efficiency is 30%
- ❖ With an intermediate state, the PV efficiency could be 60%
- ❖ One proposed material ZnTe:O
 - Is there really a gap?
 - Is it optically forbidden?
- ❖ LS3DF calculation for 3500 atom 3% O alloy [one hour on 17,000 cores of Franklin]
- ❖ Yes, there is a gap, and O induced states are very localized.



- 1) L.W. Wang, B. Lee, H. Shan, Z. Zhao, J. Meza, E. Strohmaier, D. Bailey, Proc. 2008 ACM/IEEE Conf. Super Computing, Article 65 (2008)
- 2) B. Lee, and L.W. Wang, Appl. Phys. Lett. 96, 071903 (2010)



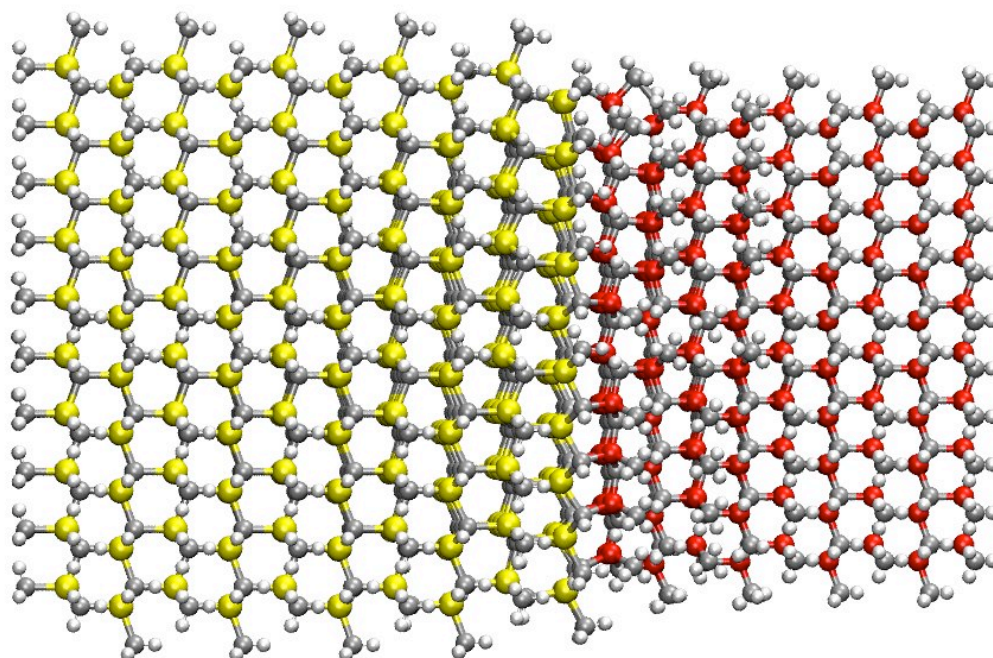
Electron and hole localization in CdSe/CdS core/shell nanorods



Isosurface of the wave function square of the CBM (green) and the VBM (red) states of the four CdS nanorods with/without CdSe core. Where (a) and (b) are for the CdSe/CdS core/shell nanorods with the Cd terminated and the Cd+S terminated surfaces, respectively, while (c) and (d) are for the pure CdS nanorods with the Cd terminated and the Cd+S terminated surfaces, respectively.

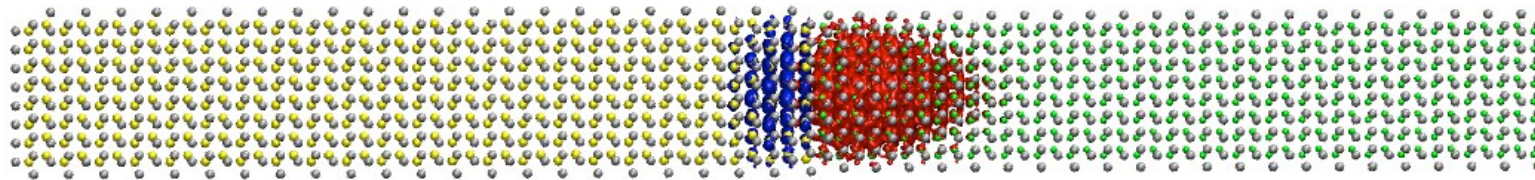
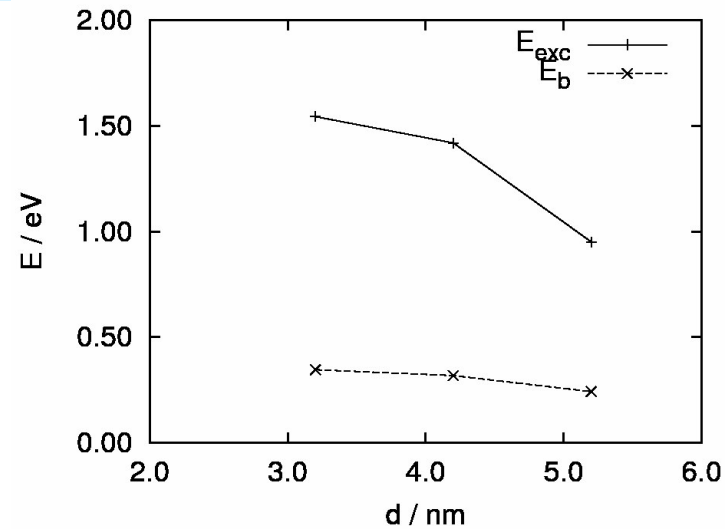
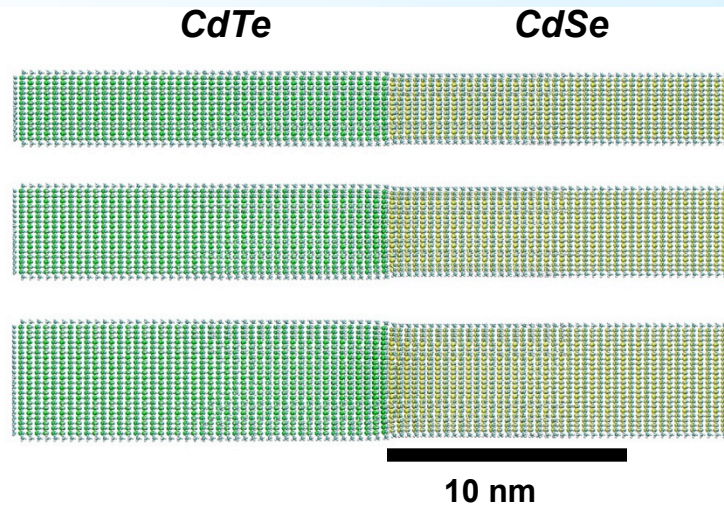
Z. Zhao, J.C. Meza, B. Lee, H. Shan, E. Strohmaier, D.H. Bailey, L.W. Wang, J. Phys: Conf. Series, 180, 012079 (2009).

ZnO/ZnS Nanorods



- ❖ Zn-surface termination, pseudo-H passivation
- ❖ LS3DF self-consistent calculation
- ❖ LDA+C band gap correction
- ❖ FSM GW+BSE approximated calculations

CdSe/CdTe Nanorods



- ❖ Exciton binding energy: 0.4 to 0.25 eV
- ❖ Exciton radiative recombination life time: $\sim 1 \mu s$
- ❖ Correlation effect: $< 1 \text{ meV}$ for ground exciton state

Slide source: Shuzhi Wang, CRD at LBL

Summary of LS3DF

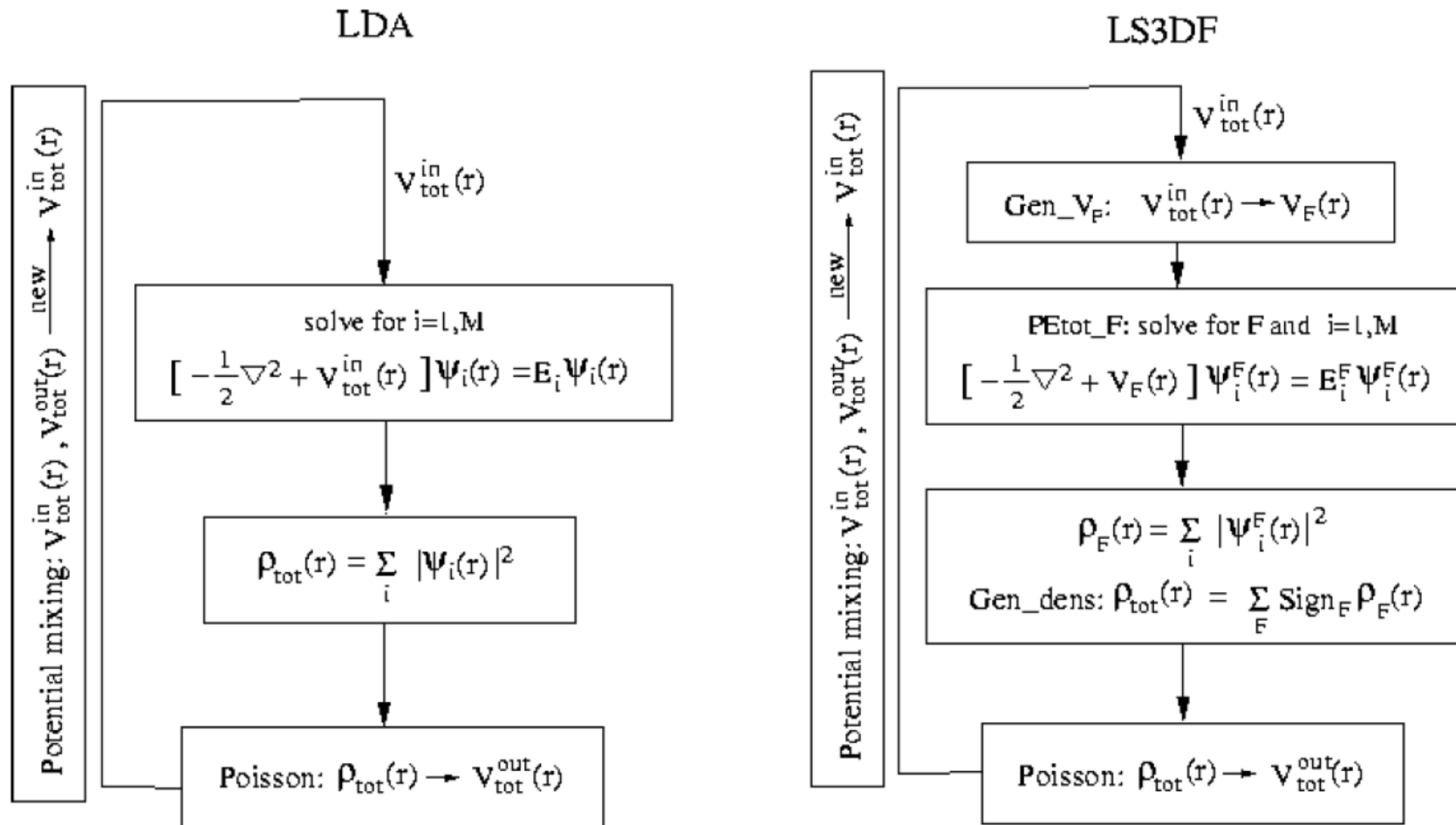
- ❖ LS3DF scales linearly to over 163,480 processors on BG/P.
- ❖ The numerical results are the same as a direct DFT based on an $O(N^3)$ algorithm, but at only $O(N)$ computational cost
- ❖ LS3DF can be used to compute electronic structures for $>10,000$ atom systems with total energy and forces in 1-2 hours. It can be 1000 times faster than $O(N^3)$ direct DFT calculations.
- ❖ Enables us to yield new scientific results predicting the efficiency of proposed new solar cell materials



More details about LS3DF code



Flow chart for LS3DF method

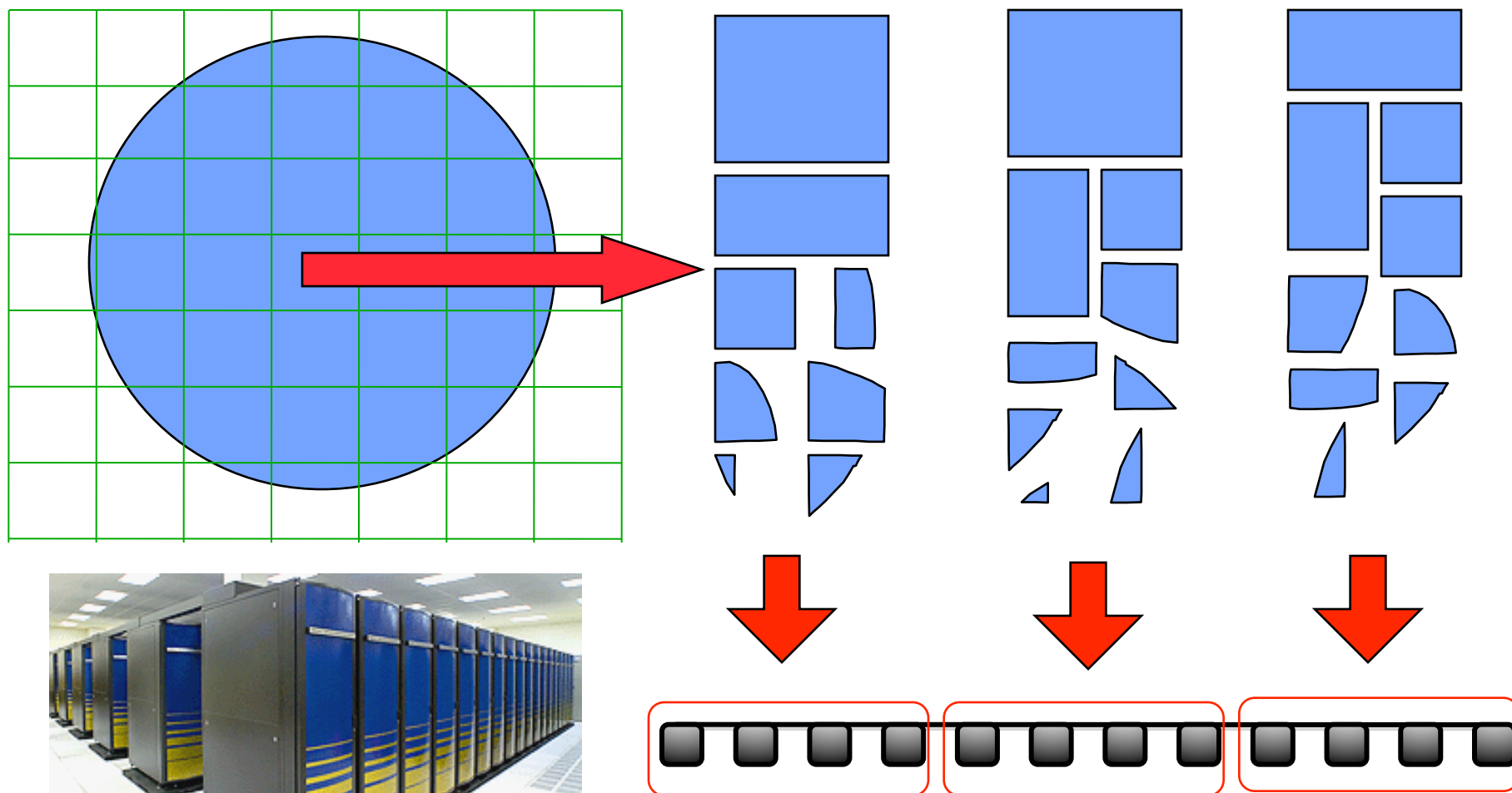


Based on PEtot code: <http://hpcrd.lbl.gov/~linwang/PEtot/PEtot.html>

ScicomP 16, May 10-14, 2010, San Francisco, CA



Schematic for LS3DF calculation



Computational overview of LS3DF code

- ❖ Variational formalism, sound mathematics
- ❖ Most time consuming part of LS3DF calculation is for the fragment wavefunctions
 - Modified from the stand alone PEtot code
 - Uses planewave pseudopotential (like VASP, Qbox)
 - All-band algorithm takes advantage of BLAS3
- ❖ 2-level parallelization:
 - q-space (Fourier space)
 - band index (i in $\psi_i(r)$)
- ❖ PEtot efficiency > 50% for large systems (e.g, more than 500 atoms), 30-40% for our fragments.



PEtot code: <http://hpcrd.lbl.gov/~linwang/PEtot/PEtot.html>

ScicomP 16, May 10-14, 2010, San Francisco, CA



Computational details

- ❖ The division into fragments is done automatically, based on atom's spatial locations
- ❖ Typical large fragments (2x2x2) have ~100 atoms and the small fragments (1x1x1) have ~ 20 atoms
- ❖ Processors are divided into M groups, each with N_p processors.
 - N_p is usually set to 16 – 128 cores
 - M is between 100 and 10,000
- ❖ Each processor group is assigned N_f fragments, according to estimated computing times, load balance within 10%.
 - N_f is typically between 8 and 100



Problems specific to BG/P

- ❖ Limited memory
- ❖ mpi_comm_split
- ❖ Core file issue
- ❖ Node mapping

```
if(inode_all.le.nnodes_tot_G) then
```

```
my_color_G=0
```

```
inode_tot_G=inode_all-1
```

```
else
```

```
my_color_G=1 ! not to be used
```

```
inode_tot_G=inode_all-nnodes_tot_G-1
```

```
endif
```

```
call mpi_comm_split(MPI_COMM_WORLD,my_color_G,  
& inode_tot_G,MY_COMM_WORLD_G,ierr)
```

```
nnodes_k_G=nnodes_tot_G
```

```
num_group_G=1
```

```
icolor_G=inode_tot_G/nnodes_k_G
```

```
ikey_G=inode_tot_G-icolor_G*nnodes_k_G
```

```
if(my_color_G.eq.0) then ! only split this part
```

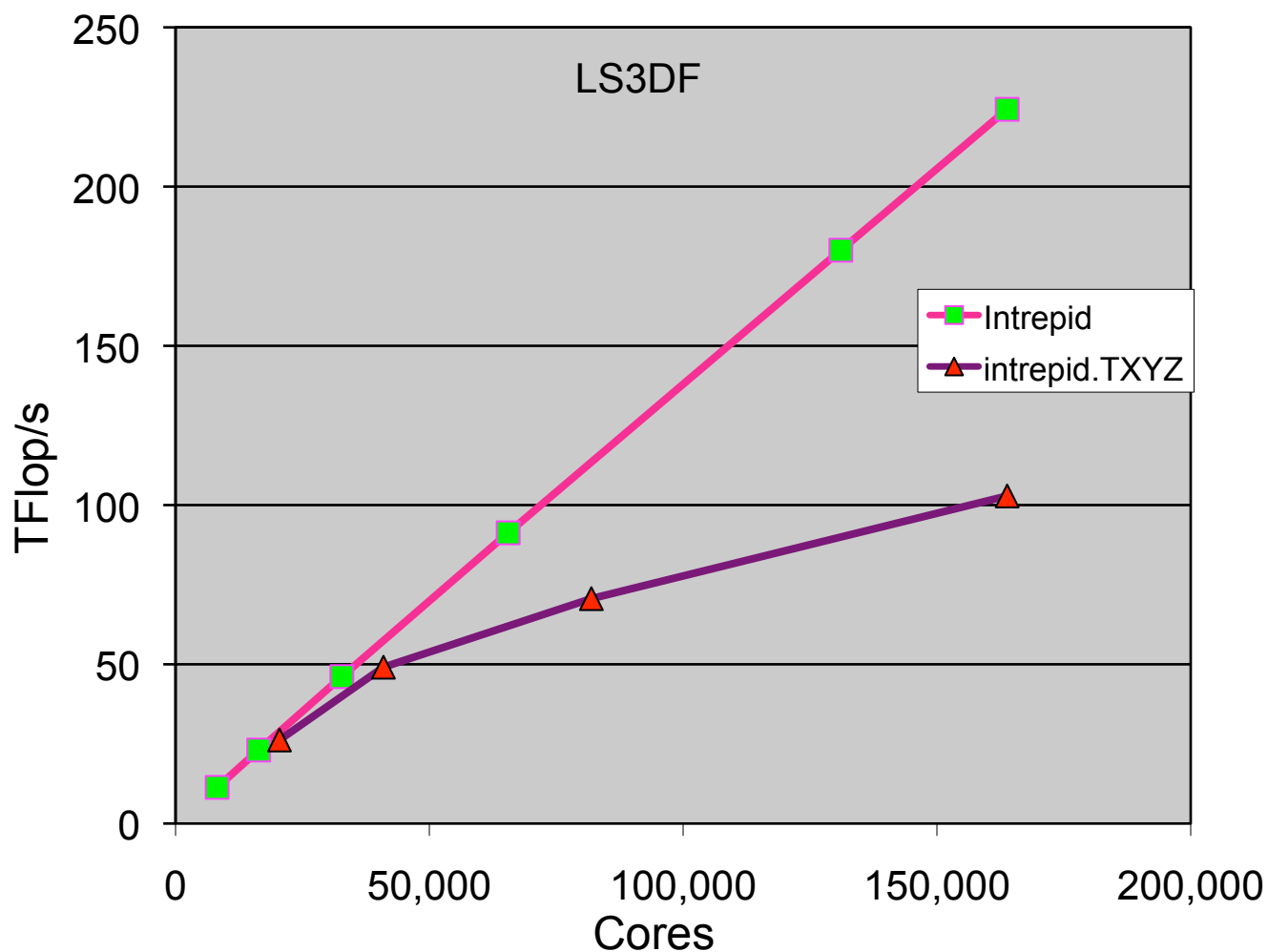
```
call mpi_comm_split(MY_COMM_WORLD_G,icolor_G,ikey_G,  
& MPI_COMM_K_G,ierr)
```

```
call mpi_comm_split(MY_COMM_WORLD_G,ikey_G,icolor_G,  
& MPI_COMM_N_G,ierr)
```

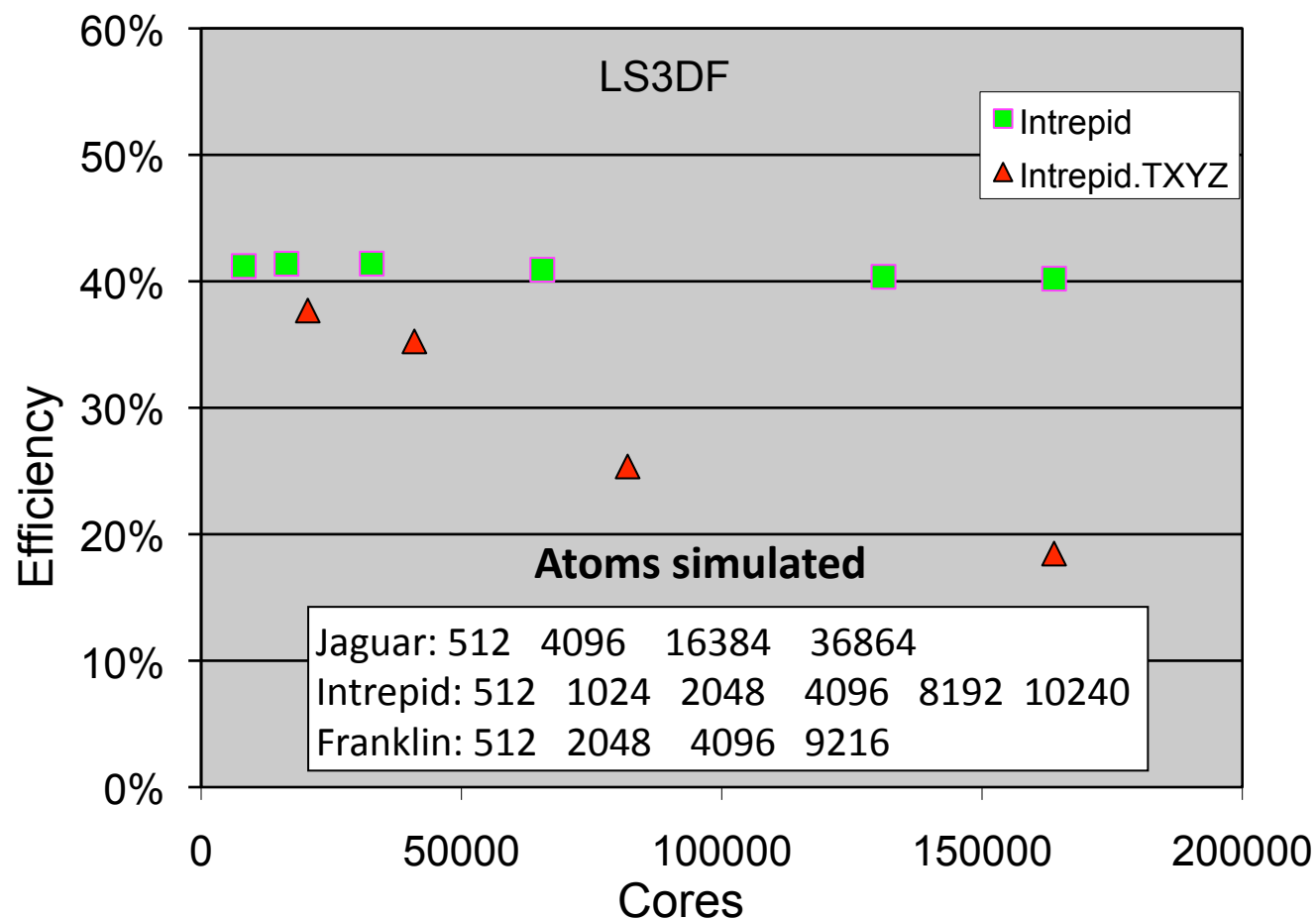
```
endif
```



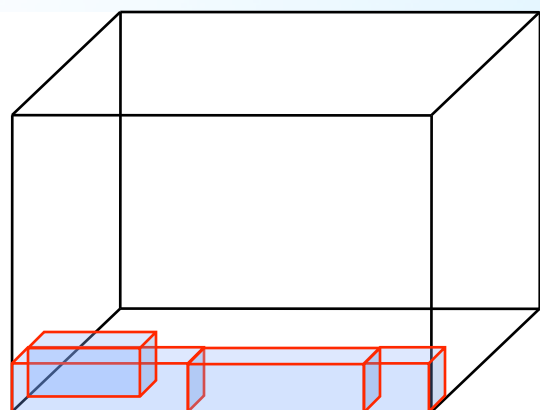
Weak scaling comparison of two mappings



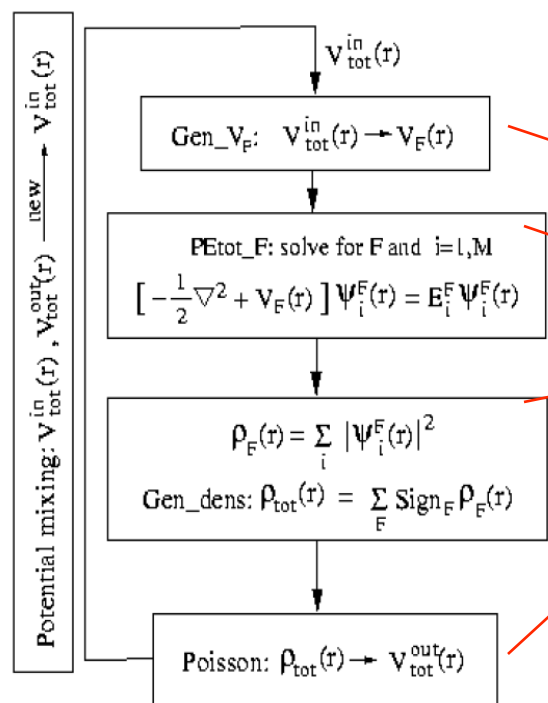
Weak scaling comparison of two mappings



The TXYZ mapping



LS3DF



For the 40 rack, the partition dimension is 40, 32, 32



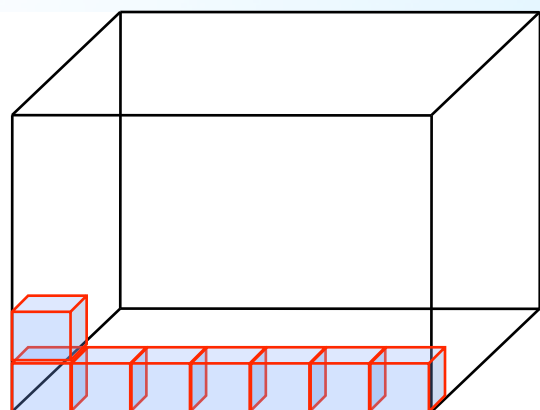
Each processor group:
32X1x1 nodes, 128 cores

Times on diff. parts of the code (sec)

cores	8192	81920	163,840
atom	512	5120	10,240
gen_VF	0.05	0.16	0.23
PEtot_F	69.30	119.53	153.98
gen_dens	0.06	0.39	0.39
Poisson	0.09	0.69	0.69

The main time-consuming part of LS3DF doesn't scale for the TXYZ mapping

The customized compact node mapping



LS3DF

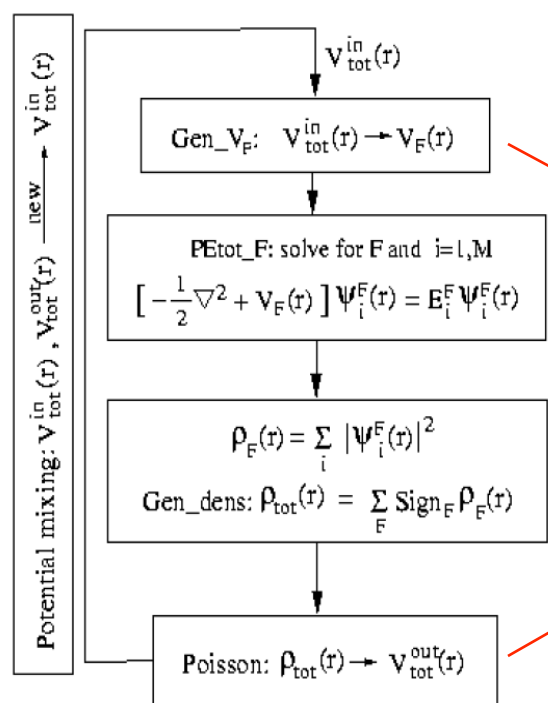
Map all the groups into identical compact cubes for good intra-group FFT communication, and inter-group load balance.



Each processor group:
4X4x2 nodes, 128 cores

Times on diff. parts of the code (sec)

cores	8,192	32,768	163,840
atom	512	2048	10,240
gen_VF	0.08	0.08	0.23
PEtot_F	69.30	68.81	69.87
gen_dens	0.08	0.14	0.37
Poisson	0.12	0.22	0.76



Perfect weak scaling

Summary about BG/P

- ❖ Node mapping was very important for LS3DF code to reach its perfect linear scaling (weak). Exploration of different node mappings when running at large concurrency should be encouraged, especially for the codes that have some spatial localities (eg., divide-and-conquer type of codes).
- ❖ Pay attention to the memory usage of MPI routines, which could become a memory bottleneck at very large scale.
- ❖ Talk to support staff who know the machine well



Acknowledgements

- ❖ Argonne Leadership Computing Facility (ALCF)
(Katherine M Riley, William Scullin, Vitali A. Morozov)
- ❖ Innovative and Novel Computational Impact on Theory and Experiment (INCITE)
- ❖ Byounggak Lee, Hongzhang Shan, Juan Meza, Eric Stroheimer, and David Bailey
- ❖ National Energy Scientific Computing Center (NERSC)
- ❖ National Center for Computational Sciences (NCCS)
(Jeff Larkin at Cray Inc)



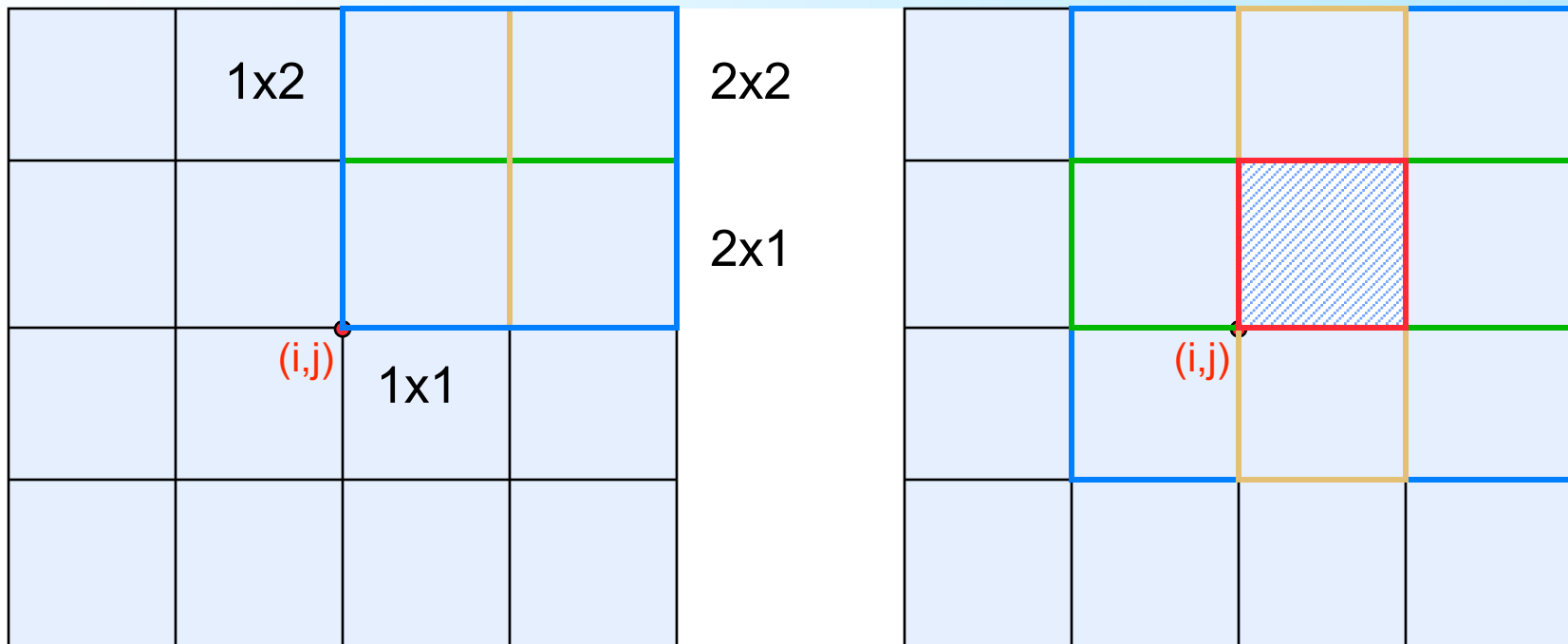
Backup Slides



ScicomP 16, San Francisco, CA, May 10-14, 2010



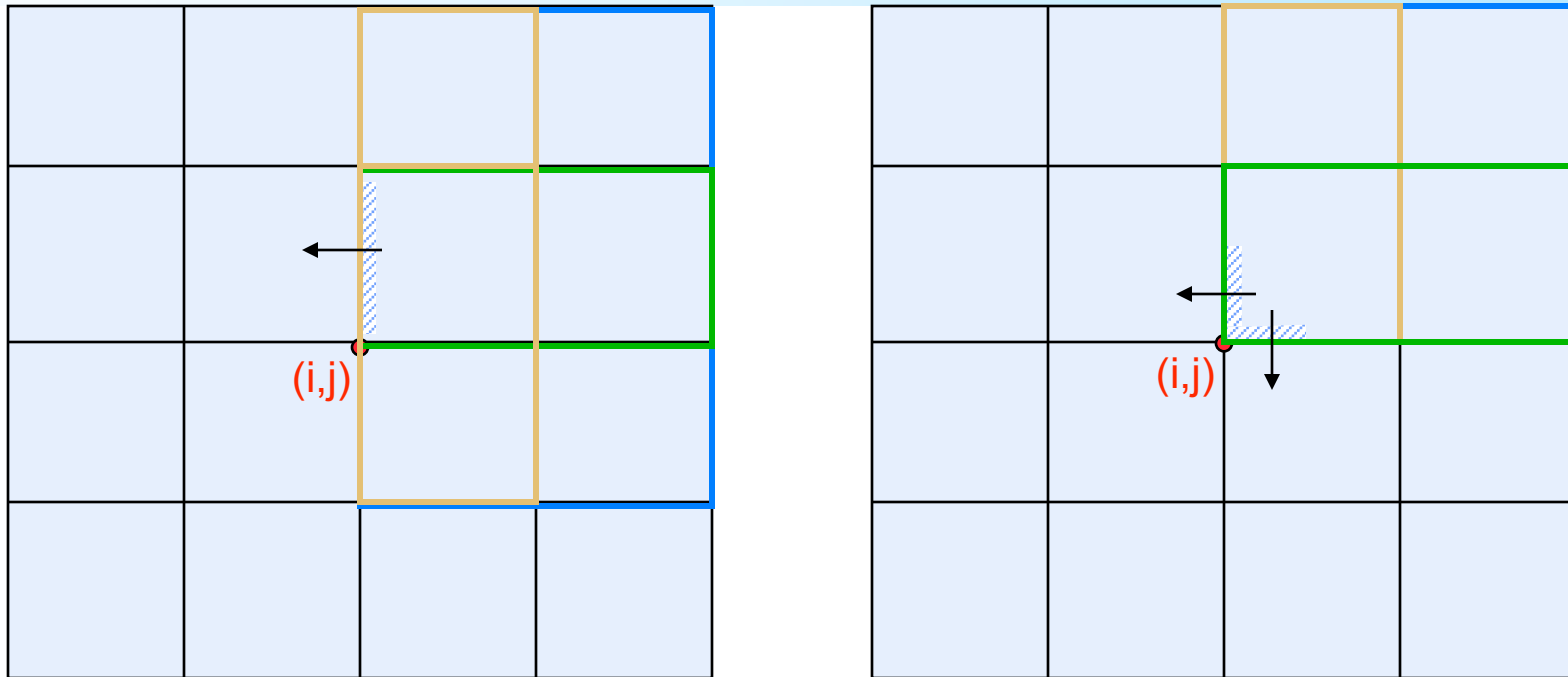
LS3DF patching scheme: 2D Example



$$\text{Total} = \sum_{(i,j)} \left\{ \begin{array}{c} \text{Blue box} \\ \text{Orange box} \\ \text{Green box} \\ \text{Red box} \end{array} - \text{Orange box} - \text{Green box} + \text{Red box} \right\}$$

Boundary effects are (nearly) cancelled out

LS3DF patching scheme: 2D example



Patching scheme is similar for 3D:

$$System = \sum_{i,j,k} \{F_{222} + F_{211} + F_{121} + F_{112} - F_{221} - F_{212} - F_{122} - F_{111}\}$$



1. L.W. Wang, Z. Zhao, and J.C. Meza, Phys. Rev. B 77, 165113 (2008).
2. Z. Zhao, J.C. Meza, L.W. Wang, J. Phys: Cond. Matt. 20, 294203 (2008).



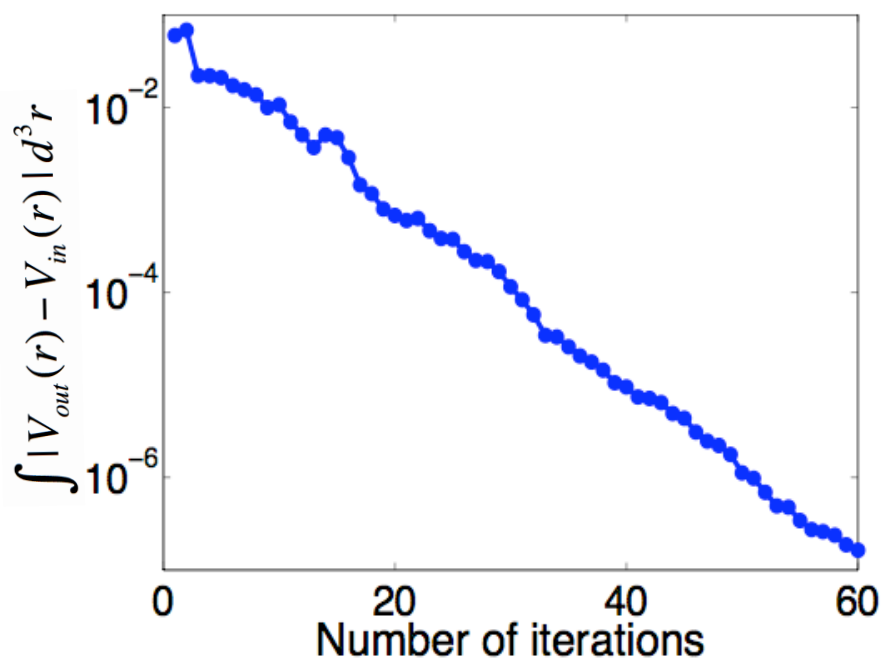
LS3DF Accuracy is determined by fragment size

- ❖ A comparison to direct LDA calculation, with an 8 atom 1x1x1 fragment size division:
 - The total energy error: 3 meV/atom \sim 0.1 kcal/mol
 - Charge density difference: 0.2%
 - Better than other numerical uncertainties (e.g. PW cut off, pseudopotential)
- ❖ Atomic force difference: 10^{-5} a.u.
 - Smaller than the typical stopping criterion for atomic relaxation
- ❖ Other properties:
 - The dipole moment error: 1.3×10^{-3} Debye/atom, 5%
 - Smaller than other numerical errors

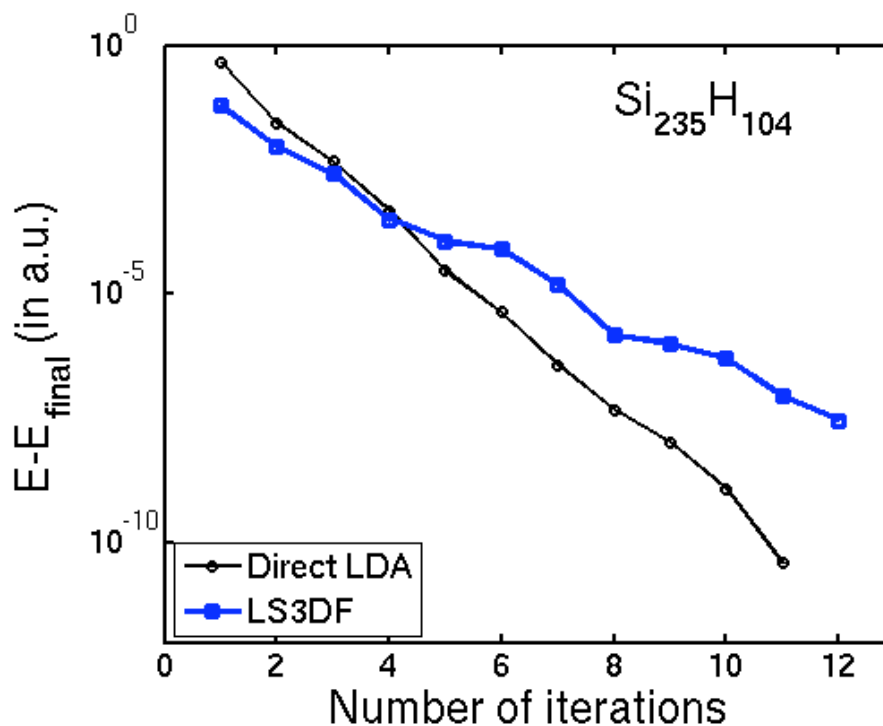
LS3DF gives essentially the same results as direct LDA



Selfconsistent convergence of LS3DF



Measured by potential

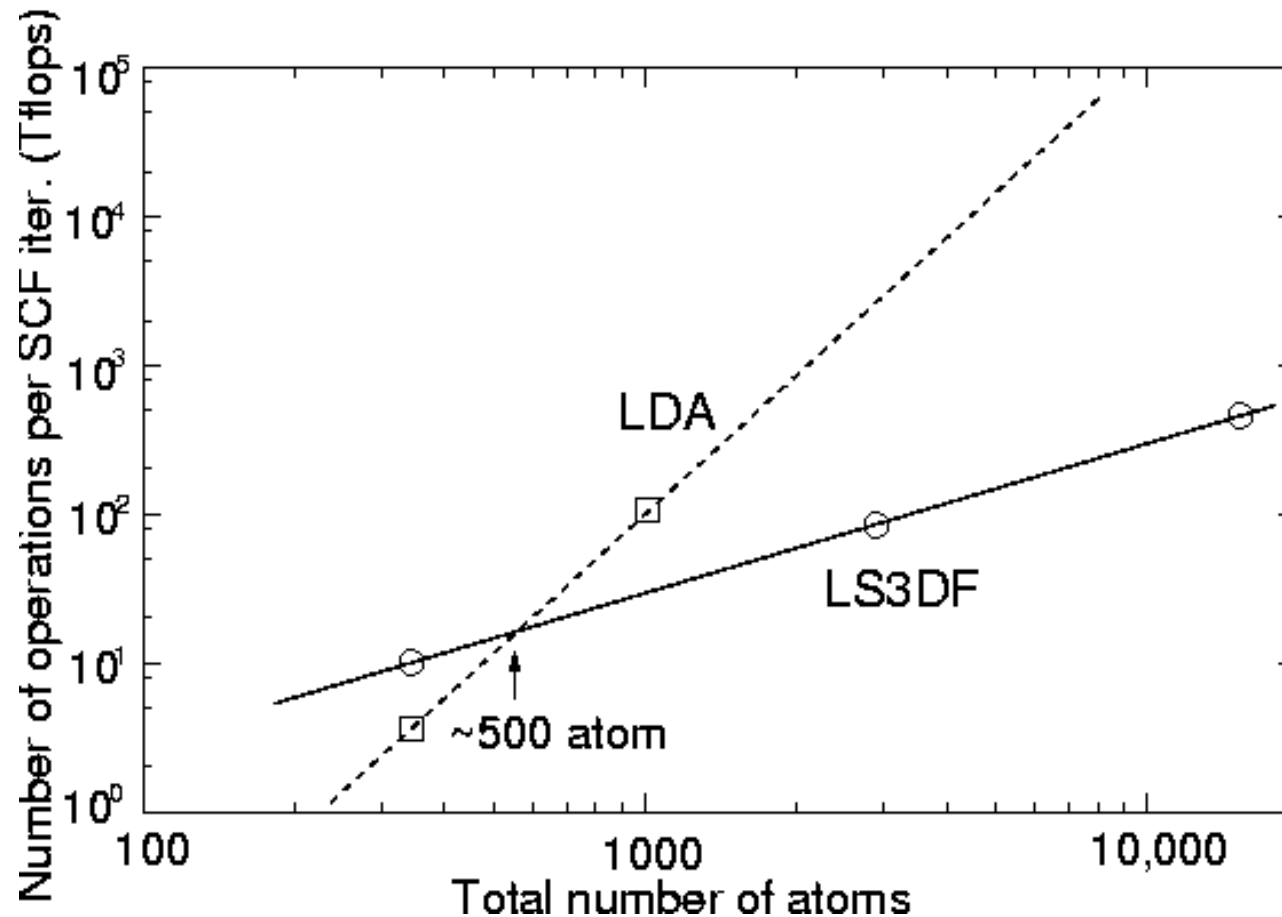


Measured by total energy

- ❖ SCF convergence of LS3DF is similar to direct LDA method
- ❖ It doesn't have the SCF problem some other $O(N)$ methods have



Operation counts



- ❖ Cross over with direct LDA method [PEtot] is 500 atoms.
- ❖ Similar to other $O(N)$ methods.

